8-31-2005

Science assessment of deaf students: considerations and implications of state accountability measurements

Julie Mountain

Science Assessment of Deaf Students:

Considerations and Implications of State Accountability Measurements

MSSE Master's Project

Submitted to the Faculty of the
Master of Science Program in Secondary Education of
Students who are Deaf or Hard of Hearing

National Technical Institute for the Deaf
ROCHESTER INSTITUTE OF TECHNOLOGY

By

Julie Ann Mountain

In partial fulfillment of the Requirements
For the Degree of Masters of Science

Rochester, NY June 21, 2005

Approved:

Dr. Jeff Porter, Project Advisor

Dr. Jerry Berent, Project Advisor

Dr. Gerald Bateman, MSSE Program Director

Abstract

As states increasingly require students to take standardized tests, it's crucial to examine the ability of such tests to accurately reflect actual student knowledge and skill within content domains. This project reviews the literature about assessment in general and written text features problematic to deaf readers. It seeks to analyze the content breadth and depth as well as the text readability and other linguistic features of Massachusetts' and New York's high school physics tests. Finally, it considers the use of tests with their potential linguistic bias for making decisions about student content mastery and potentially other high stakes decisions such as the awarding of diplomas.

Table of Contents and Figures

Title and Approval Page i

Abstract ii

Acknowledgements iii

Rationale 1

Literature Review. 4

Method 25

Results. 31

Figure 1. Item Type: Manner of Responding. 34

Figure 2. Conceptual Depth. 35

Figure 3. Conceptual Breadth 35

Figure 4. Item Type: Style of Responding. 36

Figure 5. Question Words 36

Figure 6. Readability 37

Figure 7. Comprehensibility 37

Figure 8. Unfamiliar Words Discussion 38

Discussion. 39

Conclusions and Recommendations 47

References 55

Rationale

The 2001 No Child Left Behind Law, NCLB, has prompted researchers and practitioners to re-consider the expectations we hold for our students and the ways in which we expect them to demonstrate competency. Accountability demands that all students' academic content knowledge and skills be assessed; evaluation through written tests is an attempt to compare students' actual competency with society's standards. Illustrative but not definitive, tests, observations, and other collected data relating goals with performance are used to reveal characteristics of individual students or populations. Our nation's current appeal for increased assessment, an integral element of education, has increased awareness and attention to evaluation tools' advantages and limitations.

Students who are deaf, second language learners, or otherwise struggle with English literacy are at a disadvantage when assessed on academic content using written materials. Given current law and common practice, how can educators ensure that student competency is fairly assessed? This project examines the literature and reviews what is known about testing and deaf students. The breadth and depth of two states' physics tests are analyzed. Aspects of content and language are considered for each item, then compared with state science standards and known

attributes of deaf students. Finally, suggestions about how to improve test validity for the purpose of making wise decisions are proposed.

Tests as the Bridge Between Content Standards and Student Competence

The first step in assessing student achievement requires defining the domain and clarifying the behavioral objectives that will serve to illustrate content mastery. Current legislative mandates require the establishment of a clear set of standards against which student performance is to be measured, thus both Massachusetts and New York have written documents describing their goals for students' science achievement.

Secondly, an evaluation is created for the purpose of capturing a reflection of student mastery of the articulated expectations established. Massachusetts' Comprehensive Assessment System (MCAS) and New York Regents Exams pose questions and problems through which satisfactorily answering, students are deemed to have achieved the goals set forth by their state's standards.

Finally, "assessment," the process of *making decisions* based upon collected data such as evaluation results, is undertaken. Essentially, these tests attempt to describe the student's science competency. If evaluations do not appropriately examine established standards and/or do not accurately portray student ability in the given domain, assessments, the decisions made about student performance will not be valid, undermining the entire process. Evaluations exist and are used to examine deaf students' content knowledge

in science; however, these tests may not provide good information for assessing their true science knowledge and ability.

NCLB's requisite accountability demands leave states struggling to enact fair and valid assessment procedures. Publishers take many things into consideration when designing evaluation tools; however, tests are created for the "average" student and are unable to account for every variable. A policy of equal expectations for all students and the publication of all evaluation results allows for little flexibility or subjectivity. While accommodations must be provided for students with special needs, all students must demonstrate a pre-established proficiency through on-grade-level statewide assessments. The law's sole escape route is the allowance for states to provide up to 1% of their students, those deemed as having a "significant cognitive disability," with an alternative assessment (Paige, final regulations effective 2004).

Even at their best, tests are only approximations of student competence; assessment using such evaluations must be undertaken with great caution. It is essential that tests' validity be maximized through astute scrutiny and prudent review by content experts, assessment specialists, and educational professionals. To be valid reflections of learning, tests must be aligned with objectives and instruction. They must also adequately isolate the examined domain and then appropriately and reliably measure student performance therein. We must be certain that tests examine the content

competencies that we wish to assess and that other inherent factors such as a test's linguistic form do not impact the test's results. When a test lacks strong validity for its intended purpose, this assessment tool's limitations must be considered prior to making any decision. Any important actions that are initiated or denied based on invalid evaluation results must be called into question.

Assessment of individual students' content competence is being determined on the basis of tests alone. Students may be unfairly judged based upon invalid test results and/or misinterpreted evaluations. If the tests do not reflect actual content competence, yet are used to make important decisions such as the awarding of diplomas to students or distribution of funds among schools, it is essential that law-makers, education professionals, students, their parents, and the general tax paying public are aware of this discrepancy and are given the opportunity to propose alternative assessment procedures.

Literature Review

Alignment

Resnick, editor of Research Points (2003), warns that if students are to have a fair chance to "show what they know" it is essential that there be "strong alignment" between tests and their content standards. Content mastery can be described by strength in both content breadth and depth. Any domain, be it far-reaching in scope like science or narrow such as

rotational kinematics, encompasses a wide array of knowledge, a multitude of skills and many degrees of perspicacity. In addition to its vast and varied composition, domain competence is scalable, defined by a matter of degrees. Currently, it is the responsibility of states to establish and disseminate their content standards then strive to maximize alignment of their evaluation tools.

One flaw of large-scale assessment is the strong possibility for incongruence between established "content standards" and the everyday or class-based expectations apparent to the test-takers. Teachers are charged with the responsibility of instructing students in ways that promote learning, as defined via current breadth and depth standards. Districts and teachers must become familiar with such expectations and work to ensure that students are exposed to the content and processes that their respective states require (Cavanagh, 2005). The Town of Brighton, frustrated with the New York's physics exam, ended up writing its own test so as to assess its students according to what local curriculum aimed to achieve (Freile, 2003). While this is certainly one way to approach the problem, most teachers and schools modify their curriculum and instruction to ensure that it aligns with the tests.

If the assessment process is to work, and state tests are to be useful for decision-making, there must be an obvious and strong correlation between state expectations, local experiences, and state evaluations.

Nature of tests

In examining discriminative validity of standardized tests for deaf students, Buchanan asks a poignant question: "Is it reasonable to assume that my students' true levels of achievement are being reflected by the scores from these tests?" (1973, p. 47). While the tests may have strong reliability, thus a student's score is consistent, the use of such a score for inferring actual domain competence may be invalid. As states increasingly require students to take standardized tests, it's crucial to examine the ability of such tests to accurately reflect actual student knowledge and skill within content domains.

Although assessment is usually deemed the best process to describe a student's ability, measurement error and fallible human conclusions are inevitable, contributing to misjudgments and erroneous suppositions. While critical to the field of assessment, tests are inherently limited in their ability to quantify or describe true ability Orlich (2000) argues that students in the United States may be doing better than the media often projects based on incomplete or inaccurate interpretation of simplified data. Yarroch, W. L. (1991) claims that frequently student ability is overestimated by multiple-choice tests, while Moores' (2000) asserts that for deaf students, research and experience have shown such tests to underestimate deaf student ability. Schools and students may receive high marks with one assessment while performing poorly on another. Recent NCLB-related "failures" of a "model

school" in California and a "school on the rise" in Florida provide just two examples of how dramatic the variance in assessment tools or judgments made using evaluation results can be (Hoff, 2004). Orlich (2004) addresses many of the concerns brought up by the accountability model presented in the NCLB law and describes the limitations of single-test decision-making currently in use by some states for graduation purposes. He also lists the numerous professional organizations that formally recognize the need for multiple measurements for describing student performance. Randall and his colleagues from the Arizona School for the Deaf (2000) fervently assert that "to capture the true essence of one's learning" a use of multiple measures is imperative. While the legislation that requires states set high standards and assess students accordingly is well intentioned, Moores (2004) fears that in our testing fervor we may fail to "to mitigate the potential harmful effects . . . [to] millions of American children" (2004, p. 347)

Known Attributes of Deaf Students

Like their hearing peers, deaf students are expected to take state-wide tests in order to demonstrate their progress toward established standards and their mastery of content goals. Unlike the average student, however, there is frequently a lack of correspondence between the content/general knowledge of a deaf individual and his/her language/literacy skills. Given that the assessments are primarily in written form and many deaf students' reading abilities are less than their hearing peers, linguistic bias is a serious

factor to take into consideration when using such tests to assess student ability. Rudner (1978) explains how items that are disproportionably difficult for a population of students may contain bias against that group; if deaf students cannot appropriately respond or show significantly different response accuracy to an item, that item likely contains bias against deaf students. Brown et al (1983) state that "a test item is linguistically biased if a language handicapped student has mastered a course objective but is unable to demonstrate mastery because of syntactical or lexical factors" (p. 29). They further assert: "language of test items may be the determining factor in whether or not a student demonstrates mastery of course content" (p. 24). Decisions made regarding the competency of deaf students using such biased evaluations or items should be suspect.

According to Martin and Mounty (2003), item constructions that are difficult or confusing in general "present an unfair additional challenge" to deaf individuals. "There does exist the possibility that reading level of Science is influencing the performance of hearing impaired students and thus could be a factor in the test's difficulty" Trybus, Buchanan & DiFrancesca wrote more than twenty years ago (1973, p. 59-60). In examining deaf students' performance on mathematics word problems, Kelly, Lang, Mousley, and Davis (2002) indicate that "reading comprehension level is directly related to [students' word] problem-solving abilities" (p. 120). Many sources

assert that the results of a deaf student's content competency evaluation may be dramatically skewed due to language factors.

Readability and Reading-ability

The process of deriving meaning from text is complex and multi-dimensional. Various researchers have attempted to quantify "readability" though text analysis at the word, sentence, and occasionally discourse levels. After counting target structures pertaining to word familiarity, word length, number of clauses or propositions per sentence, data is entered into a formula and a single number results, usually designed to correlate with typical grade levels used in the United States. Similarly, students are often given a standardized test from which a raw score is converted into a grade-level score meant to estimate their reading ability. While readability scores and grade-level reading ability scores do not fully encapsulate the complexity of a text or a student's skill, they do provide a succinct way to rank texts, classify students, and compare texts' readability with students' reading skill.

Features strongly affecting text comprehension

In addition to "readability" features, a number of structural and linguistic test characteristics affect the ease of determining an item's meaning, or comprehensibility. If test takers are not clear on what they are to do, what the item is asking, or what the select-type answers are saying, the integrity of the test is compromised. Tests purporting to assess student competency in a content area that are written with language beyond that

which a student can comprehend concurrently assesses student ability to understand physics and English language. Martin (2005) notes some issues with multiple choice exams based upon his extensive experience with assessment and deaf individuals. Three relevant issues regarding validity of multiple-choice items for deaf students include:

- Insufficient context
- Use of idiomatic English
- Inappropriate item content

Summarizing previous findings and conducting further analysis on exam structures, Rudner (1978) and LaSasso (1999) identify linguistic elements that can be misleading for deaf students:

- conditionals (if, when)
- comparatives (greater than, less than)
- negatives (not, un-, non-, in-)
- inferentials (should, could, because, since)
- low-information pronouns (used as place holders: "it is known . . .")
- lengthy passages

Brown, Kelly, Lang, and Kennith (1983) reaffirm the significance of conditionals and comparatives (especially negative ones), and then add other elements which test designers must use cautiously:

- Relativization [students who study pass their exams]
- Complementation [He decided that the answer was wrong]
- Conjunctions which lead to unnecessarily long compound and/or complex sentences
- Complex connectives such as: nevertheless, accordingly, respectively, and moreover

Using data from Trybus and Buchanan (1973), Rudner notes that item bias is quite prevalent in standardized achievement exams and can favor either younger on-grade-level hearing readers or older below-grade-level deaf readers.

Although exam items may consist of statements that require completion or request the student to perform some operation (identify, calculate etc.), McKee and Lang (1982) focus on question formats, noting that the linguistic manner in which an inquiry is posed may affect the results and subsequent decisions made about student performance and ranking within a group. Wh-questions are more difficult for deaf students than yes/no questions according to Quigley, Wilbur, and Montanelli (1974) but are perceived as less difficult than true/false questions according to McKee and Lang (1982). Berent (2005) explains that wh-questions alter the simple subject-verb-object word order that is easiest for deaf readers. Syntactical rearrangement because of required movement of the wh-phrase to the start of the sentence creates a later conceptual hole that must be noted, interpreted, and filled by the reader. Unlike in ASL, where such movement is optional, in English it is obligatory [SPEED, WHAT? versus What is the speed?]. Comprehension of wh-questions is easier for deaf students when the wh-question word refers to the subject and occurs in the subject rather than object position, and is easier when it is part of the main, rather than an embedded clause (Berent, 2005). Wh-movement parameters complicate a deaf student's ability to comprehend a question. Deaf learners with low English proficiency exhibit difficulty with wh-questions, especially those with movement (Berent, 1996).

There are many language features that affect text comprehension. Much of the earlier test item research was done on norm-based tests such as the Stanford Achievement Test. Occasionally, advice was provided to teachers constructing classroom assessment. However, currently, our nation is concerned with content-exams specifically designed to assess student performance on state standards. Not only are these exams given to assess student mastery of a particular subject, in some cases, they are being used to determine whether or not students will be given diplomas. Invalid use of test results may have dire consequences; professionals must not allow the test designs and our governing bodies to unfairly assess our students. Thirty years ago, Trybus, Buchanan, and DiFrancesca admonished: "further analysis regarding the language level and curriculum content of [science tests] must be made before such tests' validity for hearing impaired students can be adequately determined" (1973, p. 60). If states desire to assess student science competency, exam's language requirements must not hinder such assessment.

Not only does the language on tests need to be appropriate, the content examined must align with the domain assessed. This project seeks to examine both aspects and relate them to the performance of deaf students on state-wide physics exams.

Desired Outcomes: Defining Science Mastery

Because of new federal accountability measurements, states want to document what students know about science, or more specifically, physics. In designing evaluation tools, they must consider they outcomes they desire from school science instruction, and more importantly must design a test that will accurately measure student mastery on such outcomes. Pfeiffenger et al (1991) noted that the most difficult part of designing a test is defining test specifications due to the wide variety of curriculums and changing values. The domain of science is enormous, thus determining goals and defining standards is a challenging endeavor. The following is a description of the sources and underlying considerations that contribute to our society's expectations of student science proficiency.

The American Association for Advancement of Science (1993) notes that the precise knowledge and competencies that define science competence keeps changing. Although often viewed as a lower educational priority than Language and Math, Science remains one of the core academic subjects for the majority of primary and secondary school students in the United States. As society advances, the body of information grows, and as values change, the content emphases alter. Philosophical shifts and attitudes about relevancy impact what we expect from our students. While physics courses are taken by a lower percentage of students (20%) than are chemistry (42%) and biology (90%), (Pfeiffenberger, Zolandz, and Jones, 1991), the Physics First

movement (AAPT, 2002) aims at increasing the number of students exposed to math-based studies of energy and matter. Our society expects students to be aware to certain physics concepts and to develop skills in this and other science areas; defining the extent of student knowledge is a large undertaking given the various perspectives and the large domain that it encompasses.

Descriptions of ideals and goals for student science learning can be found in publications of professional organizations (National Science Foundation, American Association of Physics Teachers etc.) and by state education departments. The American Association for the Advancement of Science in 1993 published an often-cited reference known as *Benchmarks*. The National Research Council's Science Education Standards were published in 1995 and have also long served as a model for instruction and assessment. The most recent revising of the Massachusetts Frameworks, undertaken by a panel appointed by the Board of Education, took into account both of these organizations' publications, in addition to the Third International Mathematics and Science Study (MA frameworks, 2001).

Most states have standards for both skills and knowledge, emphasizing the belief that science is both a process and a collection of information. Knowledge-based requirements vary and are often wide in scope. Citing a study by researcher Robert Marzano, Marshal (2001) notes that the average K-12 curriculum would require about 15,500 hours to teach, while our

current school arrangement only allows for 9,000 hours of instruction time. Having clearly-defined priorities is essential if we are to ensure students master the most important elements. Inquiry processes and investigative know-how are also areas of strength that science purports to instill in students. Many objectives incorporate general core competencies introduced in other domains but essential for science achievement. In some states, science content outcomes are being integrated with literacy, whereby students are expected to read and write about science in addition to performing hands-on experiments and inquiry. In addition to language, mathematics also serves as a tool for student study of science; both graphs and equations are frequently employed in the field of physics. The formal acknowledgement and assessment of skills and knowledge provide a core set of principles to guide educators and students in their quest to develop science proficiency.

In addition to national standards, a global perspective is also sometimes considered in determining academic expectations. The Trends in International Mathematics and Science Study (TIMSS) is an attempt to compare the domain knowledge and skills of students and nations (2003). According to the TIMSS both fourth and eighth grade science students in the United States scored higher than the international average. In fourth grade, our students ranked higher than 16 of the 24 participating countries; of the countries that administered the test to a representative population, only

three scored better than their US peers. United States eighth graders ranked higher than 32 of the 44 participating countries; with 6 fully represented countries scoring better than US students. The science performance of fourth graders remained the same as in 1995 but the eighth graders improved from the 1995 and 1999 administration. As a nation, we wish to remain competitive and thus desire to rank high in such international evaluations. Although TIMMS does not publish an official standards document, their evaluation items illustrate the expectations they hold for students of science and impact our priorities and expectations.

The pressure is mounting as the NCLB deadline for the science content areas approach. By the 2005-2006 school year, all states must have science standards. By 2007-2008 they must assess science at least once in grades 3-5, 6-9, and 10-12. Despite the fact that inclusion of science test results is not mandatory for measuring adequate yearly progress (Cavanagh, 2005), science remains a core content area for primary and secondary school students. Thus, given our current obsession with accountability, student competence in the sciences must be assessed.

Content Validity of Test Items

Test items are designed to shed light on a student's knowledge or skill in a particular domain. If items within evaluation tools do not appropriately compare student actual ability with expected outcomes, Yarroch (1991), avers that assessments made based on evaluation results are rendered invalid.

Their format, structure, word choices, and other factors may determine whether or not a student understands the underlying intent well enough to respond according to his/her true ability and whether or not the item actually does reflect the student's knowledge/skill of that content. Good tests, as LaSasso (1999) describes, strive to minimize the possibility of readers getting correct answers without comprehending or of failing to answer correctly despite knowing the information assessed. While this issue could be one of random measurement error, it could also be the result of systematic linguistic bias. Results from a test that examines both language and content ability, as is the case of a content area test written above a student's reading ability, would not be valid for making decisions about student content ability due to linguistic contamination within the evaluation tool.

Because of the well-documented difficulty deaf students have with English literacy, it is essential to have a reliable and valid way of determining the linguistic demands of a given test. Furthermore, it is important to discover how systematic linguistic bias in tests can be minimized. There is much controversy over ascribing value(s) to text difficulty; however, the construct of "readability" is one attempt to reduce, into a single value, the numerous complexities that influence a reader's ability to decode and comprehend written discourse.

Comparison of student reading ability and test readability can shed light on the possible impact of language contamination within a written examination. One of the more respected formulas was created by J. Chal and E. Dale (1995) over fifty years ago; despite criticism, it continues to enjoy wide use for a variety of functions and with diverse populations. Although the formula has strong validity as referenced to other measures of reading achievements and to professional judgments as well as having predictive capabilities (0.92) as measured against other more complex or subjective standards, the developers themselves acknowledge that it accounts for only 80% of the difficulty factors identified in readability research (1995).

Like many readability formulas, the Dale-Chal formula considers the number and difficulty of words within a passage. Based on the generalization that longer sentences are usually more complex and challenging to read, passages with a high word-to-sentence ratio were considered more difficult than those composed of sentences each containing few words. Additionally, these researchers asserted that word difficulty could be measured by familiarity. Words students were unable to identify, such as those encountered infrequently or difficult to decode, were deemed "unfamiliar." In 1948, Dale and Chal compiled a list of words "familiar" to 80% of fourth graders; this vocabulary list was updated in 1981 after extensive research with thousands of school children. Thus by submitting the word-to-sentence

ratio and counts of unfamiliar words into a formula, a single numerical value indicating "readability" could be attained.

Building on the work of these and other researchers, Homan and Hewitt (1994) observed the need for an evaluation procedure specifically designed for individual isolated test items that are removed from context. They noted that validity of decisions based on test-taker scores could be impacted if the readability of test items continued to be treated as random error rather than as systematic linguistic bias. By developing their formula, they desired to systematically address this problem and validate a procedure that examined single-sentence items with short-answer choices. Their results show a definite relationship between student response to test items and the item's readability, with student accuracy on identical content material decreasing as the item became increasingly difficult to read, as indicated by their formula.

Additional readability formulas have relative strengths and weaknesses depending on the examined text style or relative difficulty. According to the Center for Cognitive Science and Educational Practice (2005), Flesch-Kincaid grade level formula is one of the most common used by educational practitioners. Flesch reading ease can be determined using a macro packaged with Microsoft Word and a score using the Kincaid formula can be found by submitting text to Readability.com (2005); these two formulas however, produce raw data rather than grade-referenced scores.

Coleman-Lieu, Fog, and SMOG formulas all convert scores to typical school grades. Their respective formulas are discussed further in the procedure section.

While readability scores may provide one piece of information about the difficulty a reader will have, it is recommended that additional features be considered before determining relative text complexity. Readability only considers the text, not the reader; within and across populations words and structures may have different meanings (Murphy, 1996). Factors beyond those that can be described though readability can impact a student's ability to comprehend text.

Israelite (1988a) details ways in which text compressibility appears to be strongly influenced by text cohesion. Data suggests that texts with cohesive features are more readily understood by deaf readers than those where connections are less explicit, even if readability scores indicate the former should be more challenging (1988b). Ewoldt (1983) makes a similar claim in theorizing that "readers will be better able to process print at the semantic level if [text has] redundancy and contextual clues" (p. 6). Cohesive ties help to make relationships, such as cause-effect, part-whole, or object-function, clear and patent. While limited in its scope of specific text characteristics conducive to comprehension, her article emphasizes the need for text to be naturally written rather than artificially constructed to satisfy a readability index or other simplification framework. Sink (2001) in examining

statewide tests, also notes the significance of explicit coherence cues in helping students make sense of text.

There are an infinite number of linguistic features that can be examined on any given test, yet little work appears to have been done to determine the characteristics that most strongly influence comprehensibility. Researchers in the fields of linguistics and deaf education have identified structures and constructs which appear to be problematic for deaf readers, and thus could contribute to a discrepancy between their ability to comprehend a test item and thus demonstrate their competence, and their true competence in the assessed domain.

Readability Alteration & Other Modifications

Despite the logic behind the idea that reducing text readability scores and/or linguistic features on an exam for students with language difficulties will lead to increased test scores, research has proved inconclusive. Murphy (1996) notes that often when an attempt is made to reduce an item's reading burden, ambiguity in its stem (item's question or fill-in portion) is created. Differences in word recognition and structural comprehension can alter the task and perception of a solution by students creating a strong source of variability in student response (Murphy, 1996). Projects that considered features in addition to readability did result in systematic, consistent data indicative of an inverse relationship between linguistic difficulty and student performance.

Bornstein and Kannapell (1971) performed the third in a series of research projects examining the possibility of linguistic bias within tests. Proclaiming "language used in multiple choice achievement test items should be no more complex than is necessary to test the examinee's knowledge of the subject matter" (p.575), they went on to assert that linguistic complexity beyond such a minimum should be regarded as a verbal overload which introduces bias. With social studies exams as their test medium, they hypothesized that simplifying the language would result in a higher mean score especially for students with limited reading ability. Using trial and error procedures to simplify the language identified as containing overly difficult vocabulary and/or unnecessarily complex syntax, the researchers then compared student performance on original and simplified items. While neither a homogeneous group of deaf high school students nor a hearing group of high school students with mean reading comprehension at the 30th percentile showed significant benefit from such modification, a group of preparatory students at Gallaudet University did perform better on the simplified language test than on the original form. Theorized reasons for the populations' different results include the overall homogeneity in aptitude and achievement of the older deaf students, the higher percentage of test completion by that group, and the nature of social studies as a strong language-dependent domain. The researchers suggested that perhaps the linguistic load only minimally contributed to test score variance or that

linguistic simplification benefited only students at or above a level of proficiency such as that exhibited by the preparatory students at Gallaudet. Similarly, in a study on mathematics word problems, Mitchell and Young (2004) found that readability alone did not account for score discrepancies between hearing and Deaf students. Nevertheless, performance differences between hearing and deaf students nevertheless persist and language appears to be related to this discrepancy.

Rivera (2003) reviewed the known studies on linguistic simplification for English Language Learners (ELLs) and concluded that more research was needed to truly understand ways in which tests can be modified and/or accommodations provided that maintain and/or ensure validity in making assessment decisions for students with English literacy skills below that of their peers. Using 4th and 6th grade science state assessment, the readability was systematically reduced on one third of the field-test items and the performances of all students on the original and linguistically simplified items compared. Monolingual students did no better on the simplified versions; the ELL student population was too small to provide statistical power to any drawn conclusions. These inconclusive results resonate with four studies performed by Abedi and colleagues between 1997 and 2001. While asserting that simplification could benefit all students and should be used (2001), they also found few to no significant differences between performance on exam versions with simplified language and provided

glossaries. One study showed a preference by ELLs for some of the modified versions; however there was no statistical difference in performance. Based on similar exam materials in which only 34% of simplified math items produced significant performance differences, researchers nevertheless concluded that both Limited English Proficient (LEP) and Fully English Proficient (FEP) students did benefit from items rewritten by content experts in linguistics and mathematics. Comparing exam accommodations for LEPs, it was found that linguistic simplification provided help to both fully LEPs and FEPs but assisted the former more, resulting in a narrowing of the performance gap between the two groups. Again, it was noted that the support provided by linguistic simplification was minimal and did not lead to significant score improvements. It was noted that performance results were inconsistent thus not supportive of linguistic simplification efficacy. Albedi, Lord and Plummer note that the "improvement between revised and original editions was small and unimpressive (1997, p.17).

For the purpose of validating a readability measurement tool, Homan and Hewitt (1994) conducted investigations of students' ability to decode and comprehend the language found on single-item tests. To assess 2nd-5th grade students' learning of a social studies curriculum (concepts taught in grades 1 through 5), multiple-choice tests were created for each grade level. Each grade level exam had 12 questions on each of seven readability levels for a total of 84 items. Only results from students scoring at or above 75% on the

content test (base-line for mastery) and/or possessing on-grade level reading skills were included in the data. As predicted, as readability increased, students were less capable at demonstrating their understanding of the concepts. The researchers concluded "there are significant differences between students' responses to test items estimated to be written on their grade readability and to test items estimated to be written above heir grade readability level" (p. 356).

Despite the confidence that researchers have in concluding that tests written above students' reading ability depress their exam scores and reduce the correlation between student test performance and actual content knowledge, efforts to reduce the linguistic burden on exams and achieve more satisfactory approximations of student true skills have not yet resulted in conclusive evidence that such attempts are beneficial. Furthermore guidelines for making such adjustments and suggestions regarding how to go about modification have yet to be realized, validated, and published.

Method

Having reviewed the literature about testing in general, science expectations, and attributes of deaf students, a review of two state's physics tests was undertaken for the purpose of discovering content and language elements and comparing them with expectations and student attributes.

Research Questions

1) What do state physics tests demand from students in terms of content/topic knowledge and complexity/depth of understanding and how do these reflect state established expectations?

2) What language level and features are used in state physics tests?

3) What do we know about deaf students' reading comprehension and its affect on their test performance?

These three questions serve as the medium for answering the driving force behind this research. State tests are required by law and are used for a variety of assessment purposes. The majority of these formal evaluations of science competence presume English literacy that is on-grade-level. Given that deaf students have been shown to lag behind their peers in reading skills, how can we maximize the validity of decisions made regarding deaf students' mastery of content through use of these language-based standardized state tests?

Research and Analysis of Tests

For this study, two fundamental text features were examined: content and language. Content analysis was done by ascribing a breadth category and a depth value to each item. Science is a broad field of knowledge and skill; physics is a bit more narrow, but still multifaceted. Test items were determined to assess one or more sub-topics. Domain knowledge may also vary in sophistication. A student may possess only surface recognition of a concept or may have an ingrained deep meaningful understanding of the concept's influence and importance. The level of thinking required to answer

test items was assessed in addition to the type of response required. Language analysis was done by examining the item formatting and the types of questions. The readability of each test as a whole was calculated using several formulas. Comprehension factors, language elements that enhance or reduce student ability to comprehend text were also counted. A variety of scores are reported comparing tests, item types, readability and other features that provide insight into the conceptual and linguistic complexities found in science tests.

It has been said that the United States' Science curriculum is "a mile wide but an inch deep." Compared with International pupils, students in the USA cover considerably more topics but primarily at a surface level (Pfeiffenberger, Zolandz, & Jones, 1991). By examining the number of objectives and the percentages of objectives within the different sub-topics, this study will report on the current state expectations for students. Each test item was classified as evaluating one or more of the following sub-topics:

- Mechanics: kinematics, dynamics
- Conservation of Energy: momentum, work, and power
- Thermodynamics: heat and temperature
- Electromagnetics
- Waves
- Radiation: electromagnetic spectrum, light, quantized energy

This study also examines the level of skill/knowledge required to correctly answer each item. Using the compact hierarchy proposed by Chase (1999), which is based on the work of Bloom (1956), each item was classified by its difficulty in terms of the level of cognitive skills entailed:

- Knowledge/Comprehension
- Application
- Higher Order Processing

An item at the lowest level simply expects students to know the "facts." An item at the middle level expects students to use information provided to solve a problem; single step calculations where the initial variables can be used directly in a provided equation are considered application type problems. However if variables need to be modified prior to using them, or multiple steps or equations are required to solve a problem, the item was classified as demanding higher order processing. This top competence level requires students to analyze, synthesize, and evaluate; it demands that students explain why or how things work and presents novel situations in which problem solving must be approached using a variety of sophisticated cognitive tools.

The tests were evaluated using a variety of readability formulas. These tools help to illustrate the linguistic sophistication of test items and provide an estimate for matching texts with students. They also indicate relative ease or difficulty for decoding and comprehending, but are generally limited to textual aspects such as vocabulary load and sentence or clausal length. While in-depth item readability analysis was abandoned in lieu of whole test evaluation, unfamiliar word counts and examples as extracted using the Living Word Vocabulary list designed by Dale and Chal in 1948,

updated during the 1980s, to include more technical and science words, and published in 1995, will be included in this study.

Exams were submitted to an on-line program (Readability.com) which is capable of calculating a variety of readability scores according to programmed data software which quickly tabulates target features such as word length, sentence length, and ratio of syllables/words or clauses/sentences. However, vocabulary was manually compared to the lexicon known by 80% of United States hearing 4th graders as identified by Dale and Chal (1995) and all unfamiliar terms were identified.

Due to the standard formatting of test items consisting of a question or statement followed by a series of answer choices, and the incongruence between formulas' expectation of connected discourse, slight punctuation modifications were made. Complete sentences that presented information, made a demand, or posed a question were left intact. Short or single word answers were made into a single sentence with commas used to separate choices. Answers consisting of both a subject and predicate were punctuated as complete sentences. Item stems which required a partial completion were punctuated with a colon at the location of the blank, and answer possibilities were inserted with commas between each choice unless answers consisted of a full clause, in which case each was punctuated as a complete sentence.

Given the mathematical nature of the exam, there were many instances of values being followed by their respective units (e.g. 10 Joules,

5N, 15m/s², 8Ω). Numbers and units were separated by spaces and were counted as two words. Most formulas considered single letter units "familiar" while noting that multi-letter units would be more difficult and thus unfamiliar. Numbers raised to a power were separated from their exponential notation; the power notation was attached to the unit resulting in a single unfamiliar word. Student fluency with reading number-units has not been evaluated by readability formula constructors and may influence both the student's approximate reading skill and the test's reading score. Because none of the readability formulas were designed with technical numerical formatting in mind, test items had to be consistently represented in a manner that would approximate their relative ease/difficulty.

The following formulas were used to calculate readability scores:

Flesch Index	= 106.835-4.6*syllables/wd-1.015*wd/sentence	= 0 (hard)-100 (easy)
Kincaid	= 11.8*syllables/wd+wd/sentences-15.9	= difficulty 5.5-16.3
Coleman-Lieu	= 5.8*characters/wd-0.3*sentences(100*words)-15.8	= grade equivalent
Fog index	= 0.4(wds/sentence + 100*((words>=3syllabols)/words))	= school grade
SMOG-grading	= square root of ((30*(words>=6syllables)/sentence)+3	= school grade

A second linguistic analysis was undertaken to examine the text from a broader perspective. Text was scrutinized for features that were closely tied to comprehension rather than simple decoding. Each item was examined for linguistic structures or elements that have been shown to cause difficulty in comprehending text and test item intent. Low inference pronouns and negation in its various forms were identified when they occurred. The number and percentage of passive sentences was noted. An in-

depth look for comparative structures was also undertaken. Finally, conditional statements were examined.

Examined Resources

Both Massachusetts and New York release their state physics exams. The items examined were taken from tests retrieved using each state's respective Department of Education web site. Examined editions were the Massachusetts Comprehensive Assessment Test "Grade 9/10 Introductory Physics Pilot" test given in the spring of 2004 and the New York Regents Exam "Physical Settings: Physics" administrated June 16, 2004. The two tests are given to High School Students and may be applied toward graduation. While neither Massachusetts nor New York explicitly requires students to pass the physics test, both states do withhold diplomas from students failing to pass required tests (Kadamus 2004). Massachusetts' students are required to take the physics exam, but currently do not need a particular score on this test for graduation. New York Students must pass one Regent's science test, but may choose the topic they would prefer: Biology, Chemistry, Physics, or Earth Science.

Results

When determining alignment between standards and evaluations, the grain size (AAAS, 1993) of a domain's included knowledge/skills must be taken into account and the emphasis of each expectation should be apparent in both the state standards and in the assessment protocol. While some content learning occurs in pieces, other understanding only comes from experiencing the bigger picture. Evaluation needs to assess student mastery

of details and the gestalt in accordance with the standards set forth by the state as constituting science mastery.

The Commonwealth of Massachusetts expects educational proficiency demonstrated in academic areas as defined by content frameworks. Most recently published in 2001, the science/technology and engineering frameworks consist of four learning standard strands, a substantial prefix (including goals for inquiry) and appendix describing additional factors crucial to the attainment of science achievement. At the high school level, standards may be viewed in two different ways: (a) separated into content area classes or (b) organized into integrated courses. For the purpose of this study, the Physics 9/10 section was used. Consisting of six broad topics, it contains 67 standards, 25 of which are "core." The Massachusetts comprehensive exam evaluates student competency on only the core standards. A document describing the alignment between standards and test items is provided, and for this edition, most core standards were reflected through one item. However several standards in the first topic, Motion and Forces, were assessed through two or three items. Overall, it appears that great care is taken to ensure and illustrate alignment between state frameworks and state assessment items. Nevertheless, efforts to keep the test to a reasonable length could have devastating implications for the assessment procedure— due to the low item-standard ratio, students have very few opportunities to demonstrate mastery of each individual state goal.

New York State expects educational proficiency in academic areas as defined by benchmarks. Most recently published in 1996, the Learning Standards for Mathematics, Science and Technology sets forth seven standards to which schools/students are to be held accountable to prior to commencement. Two standards describe expectations specific to physics. Standard 4 is divided into the physical environment (five key ideas) and the living environment (seven key ideas). Standard 5 focuses on technology with seven key ideas. A "map to core curriculum" is provided to illustrate the alignment between key ideas/standards and test items. Standards four and five have explicit physics foci, thus each text item is identified with a comprehensive objective and/or or process skill from these sections. Additionally, items are referenced to key ideas for science inquiry. Because of the strong connection between science and math, items are also referenced to math key ideas. Standard 1 has three key ideas each for math and science and an additional key idea for technology. Standard 3 has seven key ideas for math. The remaining Standards, 2, 6, and 7 are referenced minimally for this test.

The Massachusetts exam has one part for a total of 27 items. It consists of 20 multiple-choice questions and 2 open response questions (each treated as a cluster with 3 and 4 items respectively). Thus 24% of the items require students to generate answers while 76% ask students to select from four choices. [See figure 1].

The New York exam has three parts for a total of 73 points. Part A consists of 35 multiple-choice items. Part B consists of 13 multiple-choice and 11 open-response items. In part B, nine items stand alone, and the remaining fifteen of these items belong to one of 6 clusters each with 2-3 questions based on identical information or data. Part C consists of 14 open response questions that make up 5 clusters. Considering all items, students must generate their own response to 34%, whereas they select from up to four choices for the remaining 66%. [See figure 1].

Figure 1. Item Type: Manner of Responding

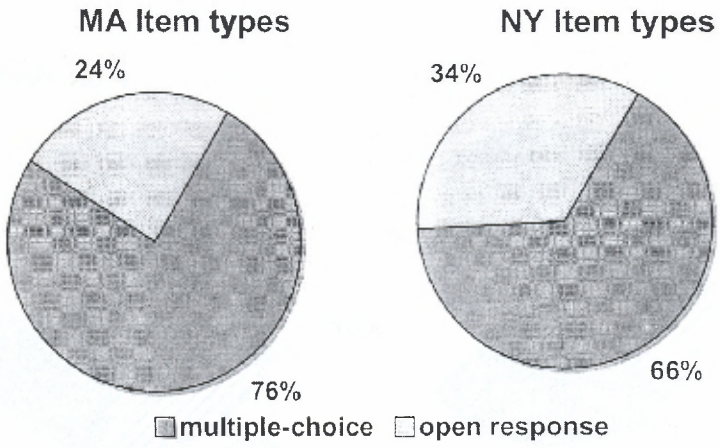

Figure 2. Conceptual Depth

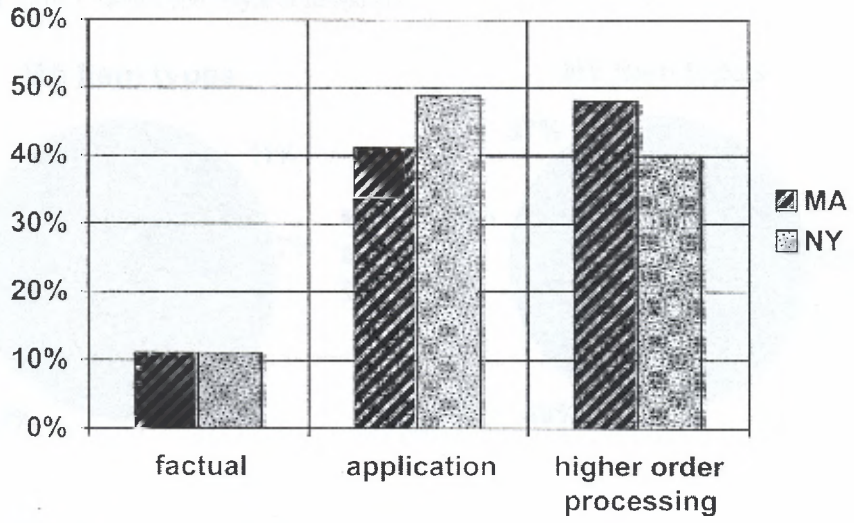

Figure 3. Conceptual Breadth

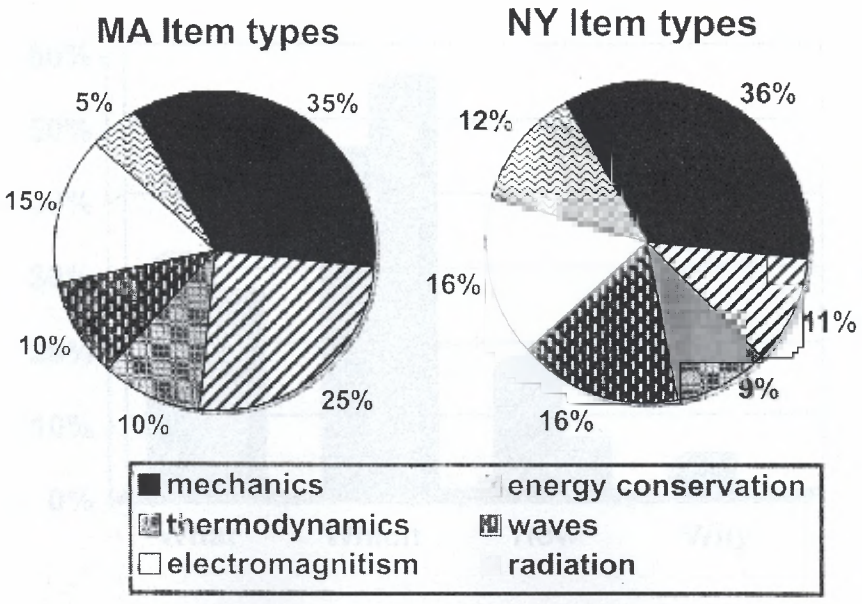

Figure 4. Item Type: Style of Response

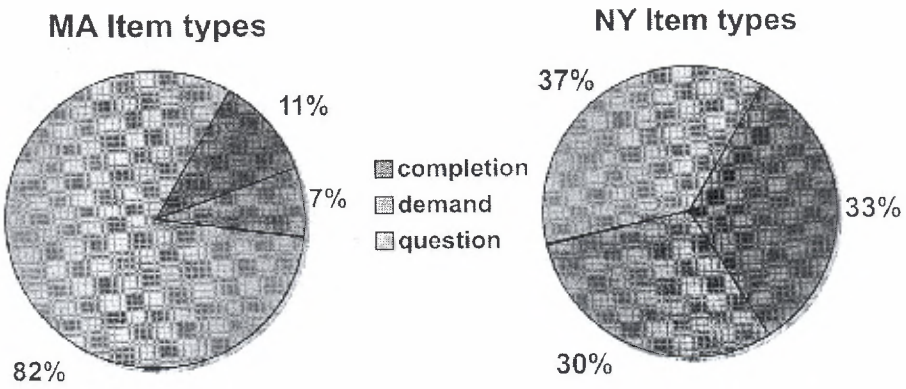

Figure 5. Question Words

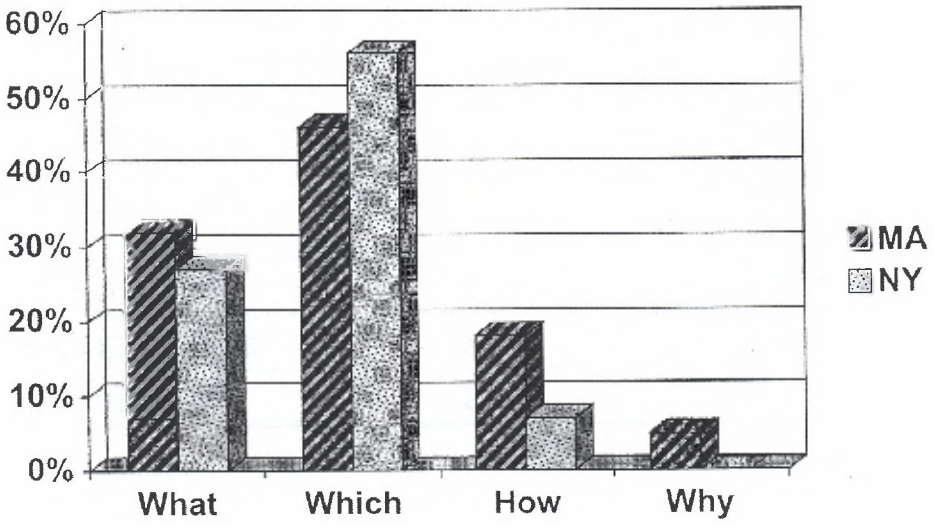

Figure 6. Readability

	MA	NY	Together
Fletch index	70	64.5	66.7
Fletch-Kincaid	7.5	9.5	8.8
Kincaid	7.1	9.1	8.4
ARI	9.9	8.8	7.9
Coleman-Liau	9.7	9.8	9.8
Fog	9.9	12.4	11.5

Figure 7. Comprehensibility

	MA	NY	Together
Low information pronouns (it)	1	2	3
Negation	1	4	5
Passives	14	44	56
Comparatives	18	25	43
Conditionals	2	5	7

Figure 8: Unfamiliar Words

Test taking and relational		Math and Science		General	
according	opposite	alpha	parallel	advantage
affect (ing)	predict	aluminum	projectile	aircraft
approaching	procedure	angle	property	applied
approximately	produce (d)	asteroid	proportional	asphalt
assume (ing)	provided	cathode	radius	barrier
calculate	relationship	circular	slope	beaker
characteristics	relative	copper	source	booklet
combined	represent (s)	core	sphere	calm
compare (d)	require (d) (s)	cross-sectional	stopwatches	emergency
compose	series	data	tide (s)	information
conclude (sions)	standard	deflect (ed)	unit	lightbulb
conduct (ed)	stationary	diameter	vertical	mission
consisting	steadily	disturbance		natural
construct (ed)	substitution	elongation		object
correspond (ing)	suggest	equations		performing
decrease (d)	support (s)	exerted		plastic
demonstrate(ion)	transform	expand		pluck
detect	various (ying)	flow		rebound
determine		function		shuttle
diagram		graph		situation
effect		grid		staircase
elapsed		hydrogen		stroller
equivalent		hypothesis		supply (ied)
evidence		incline		undergoes
external		initial		unless
factor (s)		inline		vehicle
gained		interval		versus
horizontal		laboratory		weightlifters
identical		laser		
identify		location		
including		material		
increase (d)		maximum		
indicate		minimum		
inversely		monochromatic		
label		navigation		
neglect		node/antinode		
obtain		observation		
occasionally		observer		
occur (s)		output		

Discussion

Tests and Expectations

In order for a test to be a valid and useful assessment tool, its measurement domain must be clearly defined and its items must examine content within its described domain. Both New York and Massachusetts have spent considerable energy in creating, revising, and disseminating their academic expectations to educators and have likewise been conscientious in illustrating that their evaluation items align with their respective standards.

One concern with standardized tests, and multiple-choice items in particular, is the tendency for items to primarily assess lower-level thinking skills such as factual recall and comprehension. A review of the two tests indicates that considerably more items address application and higher order processes than factual knowledge, a result that resonates well with many educators and professional organizations (AAAS, 1993). Each test contains about 10% factual or identification questions with the remaining 90% being evenly distributed between items that require students to calculate and use information or analyze, synthesize, and evaluate information, often incorporating multiple skills or elements of knowledge. [See figure 2].

Both exams also focus on the integrative nature of science with other domains of learning. Incorporating inquiry, they demand that students engage in the scientific method and assess students' ability to solve problems and interpret data. As stated in Massachusetts' frameworks document

(1996): "What is known does not stand separate from how it is known" (p. 5). As indicated by New York's Map to Key Concepts, key ideas from three different standards are integral to science mastery. Each item did, however assess content knowledge from one of six physics' areas. [See figure 3].

Tests consist of three different item types: questions, demands, and partial completion. Figure 4 shows the respective percentages for each test. The questions, are nearly all wh-questions, either asking "what" or 'which." [See figure 6].

Language

While ideal, domain isolation is nearly impossible; with written tests, language will contribute to evaluation results. Maximized validity requires minimizing the impact of language on the assessment of student content-competency. It appears that both Massachusetts and New York have made some linguistic considerations, however the items presented on their tests contain language that may invalidate test results for students with reading difficulties.

Several language elements were analyzed for the purpose of comparing the two tests' linguistic complexity and for illustrating the significant infusion of language demands on what is intended to be an evaluation of science content knowledge and skill. Language influence was examined from two related but distinct perspectives. First, "readability" was examined using a variety of different formulas created and manipulated by researchers in an

effort to quantify text difficulty. [See figure 6]. During this process, "unfamiliar" words were highlighted and classified according to their interdependence with physics, science/math, and the testing process itself. [See figure 7].

Second, features previously identified to be problematic for deaf readers and test-taking students, were examined. [See figure 8]. Neither low-information pronouns nor negatives appeared frequently, indicating either intentional exclusion of such features on the part of test designers or fortune. Passives, conditionals, and comparatives did appear with some regularity and thus illustrate a possible barrier between an item's intended purpose and the students' ability to answer according to their actual knowledge/skill. If language comprehension prevents a student from accessing an item, any inference about the student's ability on contained content will not be valid.

Readability

A variety of readability formulas were applied to the respective tests. [See figure 6]. With Kincaid providing one notable exception, New York's exam received higher readability scores, indicating its text should be more difficult to read than that found in Massachusetts' exam. Grade equivalents ranged from 8.8 to 12.4 (with a median of 9.65 and mean of 9.9) for New York and 7.1-9.9, (with a median of 9.7 and mean 8.0) for Massachusetts. Given that this exam is designed for 9th-10th grade students, such scores might at first appear reasonable. However, any student who reads even slightly below

his/her "average" peers may have difficulty accurately interpreting the items' stems and/or answers. Such difficulty may result in test scores being contaminated by assessment of language rather than functioning primarily for assessment of science content. Given that even the most literate and successful deaf students generally lag behind their hearing peers in reading competence, the high reading demand of these exams is cause for concern when the test is intended to evaluate science mastery.

Vocabulary

One of the more interesting findings in this project came from the process of identifying "unfamiliar" words within items. A word was determined to be "unfamiliar" if less than 80% of hearing fourth graders did not know it, according to the research that produced "the living word vocabulary" list (cited in Chal & Dale, 1995). Because the average deaf student graduates with a 4th grade reading level (Traxler, 2003), and even high performing deaf students usually read below that of their hearing peers (Traxler, 2000), 4th grade readability standards were ideal for this project. Figure 7 organizes many of the lexical items found on the test, distributing them into columns based on familiarity and on explicitness from instruction.

As an academic and professional field of study, physics employs numerous words that have explicit, defined meanings. Called "technical vocabulary" or "content specific terms," these words' precise meanings are usually defined in student texts, found in their glossaries, or extensively

explained during lectures. Whereas some words are used only within the field, others may have broader applications with meanings that are different or distorted. Because these terms are accepted within the field and usually necessary for communicating about the subject, it is appropriate that student understanding and application of technical vocabulary be assessed. Other, non-technical words can be divided into three categories. General science and math words are closely related to physics but may either not be explicitly defined within the context of instruction or may have assumed, implicit meanings. Other academic vocabulary such as procedural or test-taking words includes those that explain what students are to do in order to demonstrate understanding of the question. Finally, there are general miscellaneous words that may or may not be necessary to convey meaning or item intent.

Unfamiliar words within written information, questions, and response choices, can have a tremendous impact on a student's ability to comprehend an item's intent and respond appropriately. Martin and Mounty (2003) highlight the issue of words that are used in tests but not frequently found in non-testing situations. They note that unless a word itself is being tested or is necessary to convey examined content, the use of unfamiliar words in written evaluations can put weak readers at a distinct disadvantage.

Inherent in most high school courses, physics not withstanding, is an element of vocabulary instruction; for deaf students, lexical learning is

frequently a more explicit instructional process than for their hearing peers. While textbooks usually define technical words, the definitions themselves may contain words with which a deaf student may be unfamiliar. Chaing-Soon et al. (1993) examined the readability of popular high school science textbooks. Using the Fry formula, they found that about 30% of the physics texts were rated with college level readabilities. Thus teachers and students are faced with difficulty in explaining and "owning" new words when much of the supporting text consists of are unfamiliar words. Additionally, the linguistic structures necessary to define some words may be beyond the students' reading ability. Related words are sometimes taught elsewhere (e.g. math class) and transfer is expected or assumed but other terms are acquired by hearing students incidentally and through frequent use in various contexts.

A frequently cited reason for deaf students' general vocabulary poverty is their lack of access to spoken English (Marschark, Lang, & Albertini, 2002). Some students may not have encountered some of the related or general words in their everyday (self chosen) or school-related (teacher assigned) readings due to the tendency to select or be provided with texts at their below-grade-level reading ability. Students who are given on-grade-level texts may still struggle due to either their own lack of reading skills, or the difficulty in simultaneous language decoding/comprehending and content learning. Tests themselves may create a systematic "vocabulary bias" but

teachers also are responsible to ensure that students have had sufficient exposure to words likely to be used on the test and that content specific lexicon has been explicitly taught and reinforced in the classroom. While most teachers know the importance of new vocabulary instruction, those who wish to ensure a high degree of correlation between actual student ability of the content material and tests' reflection of such must be aware of all content-specific, content-related, and general-test-related words likely to appear within evaluations. Teachers who work with student populations delayed or weak in language competence must be acutely aware of prior student knowledge and endeavor to teach and/or expose students to the full range of vocabulary.

Comprehensibility

A good assessment strives to minimize language barriers to the effective communication of an item's intent and its provided potential responses. Many linguistic components can distract from the context examined within a test or item. Stems should be written to clearly indicate the type of response required and answers should consist of similarly structured responses. Items on both the Massachusetts and New York exams are clear and, when appropriate, provide ample visual support. Typical of multiple-choice tests, the stems are succinct, removed from context, and lack explicitness, thus forcing the test taker to draw inferences and make connections.

Following good practice, both states' exams make clear the relationship between referents and their antecedents, and minimize low-information pronouns such as "it" [MA= 2 NY=1]. The use of negatives is known to create confusion and should be used with caution especially when test takers' language competence may be below that of their concept competence. While directions to *neglect* (not account for) appears, genuine negatives appear in very few instances [MA =1, NY = 4]. Both Massachusetts and New York have constructed their exams around sound assessment principles, indicating conscious attention to both the structure and content of their items.

Question structures can pose difficulty for students because of by their complex syntax. Approximately half of the items contained questions: 63% of Massachusetts' items and 48% of New York's items. Of those questions, the majority asked "what" requiring the identification of a correct answer, or "which" requiring the choice between possible answers.

Passive structures also can be problematic given their deviation from the typical subject-verb-object semantics found in active sentences. The noun that would exist as the object in an active sentence becomes the subject in a passive sentence causing possible confusion about who or what is performing the action of the sentence. Passives appeared 56 times in the two exams, with 10% of Massachusetts and 30% of New York items containing one or more passive structure.

Conditionals are the typical language structures used in hypothesizing and drawing conclusions. Employing an "If. . . . then" statement or a clause connected via "unless," such sentences illustrate hypothetical situations. While use of conditional structures may be needed to convey complex concepts and to assess student proficiency thereof, the linguistic form is difficult for deaf readers to comprehend. Approximately 10% of each exam's items contain conditionals. This may be a reasonable balance between the need to include them for assessing student's ability to grasp conditional concepts and the desire to minimize linguistic difficulty.

Deaf students often struggle with relational language, thus comparatives within an item stem or answer choices may cause difficulty in the selection of a response that truly reflects a student's understanding of the examined concept. Nevertheless, science often requires students to compare variables and predict, determine or conclude the affect variables have on one another. In order to assess a student's mastery of concepts such as variable change and relationships, comparative and superlative linguistic structures must be used. Figure 7 shows the composition of such features in the examined tests.

Conclusions and recommendations

What can we do?

Assessment is an integral component of education and thus worthy of substantial contemplation. Intended to assist in the decision making process,

evaluations are powerful tools which must be wielded with caution. Assessing student performance in content areas, such as physics, through the use of written tests highlights many challenges inherent to the nature of evaluation and issues specific to deaf students.

Evaluations designed to assess student performance in a particular domain must first maximize alignment between standards and test items and must minimize content contamination by ensuring that tool format or language bias does not impact a students' ability to illustrate their knowledge in the intended area. Secondly, students need to have solid content instruction in the areas that are evaluated and may need exposure to formatting and language intrinsic to assessment of the domain. Third, an understanding and acceptance of the imperfect nature of assessment procedures requires that flexibility in decision-making be exercised by officials in judging student competence and subsequently recognizing mastery through the awarding of high school diplomas.

Maximizing Evaluation Tool Effectiveness

Tests and other evaluation measures must be carefully designed to maximize their ability to reflect true student knowledge. There must be explicit alignment between established standards and evaluated knowledge/skills. Furthermore, the breadth and depth of domain competence tested must reflect the expectations reflected in the state's frameworks or benchmarks publication.

Test writers must use prudent judgment in designing items to ensure that students will grasp the intent and not be distracted or confused by the format. Elements such as stem and answer choice construction have characteristics that affect the possibility of accurate interpretation. Question format may impact item comprehensibility and for students with language difficulties becomes a serious area of concern. Use of low information pronouns and negatives can create ambiguity and confusion and thus should be avoided or minimized in test items. While some structures are unavoidable for the purpose of evaluation, those that are unnecessarily difficult to comprehend should be eliminated or modified. Structures such as comparatives, conditionals, and passives should be carefully considered before they are included on an exam that is to be taken by Deaf or LEP students. The language of a written exam has a profound influence on how meaningfully its results reflect the test taker's competence in the evaluated domain and therefore on the validity of educators' decisions concerning the student's knowledge.

Most academic content evaluations necessitate some degree of language competence on the part of the test taker. Student performance on paper-and-pencil science exams is thus dependent on both content knowledge and reading ability. There remains a great need for research that looks into how content-curriculum exams can be better constructed to maximize the

relationship between deaf students' true science knowledge and their apparent ability as reflected by test performance.

Given the impossibility of entirely separate cognitive and linguistic domains within testing parameters, we must strive to identify and utilize effective strategies to maximize the correspondence between actual student knowledge/skills and apparent ability as reflected in exam scores.

Teaching Content and Language

A strong alignment of the evaluation with established expectations is critical if valid decisions are to be made. Marshall (2005) said it well: "a teacher can cover only a portion of the total curriculum, and the tests can assess only a portion. For students to do well, the portion that's taught needs to overlap with the portion that's tested" (p. 30). Schools and teachers must become intimately aware of these expectations and must design curriculum and instruction in ways that will ensure the greatest possibility of student achievement.

Not only will student content competence help students to be more able to respond correctly to tested material, it may help them to better understand the questions. Ewoldt (1983) emphasizes the relationship between text comprehension and familiarity with the topic, implying that of two equally skilled readers, the one with more science experience is likely to have greater ease with science text. Ability to draw inferences also appears to be

enhanced by students' general knowledge of the topic and familiarity with content specific vocabulary (Sink 2001).

Teachers must endeavor to promote at least the minimum language skills necessary for using a written content evaluation. Students need to understand the process of selecting "the best possible answer" and must know the meaning of demand words such as "identify," "describe," "graph," and calculate." Skills particularly important in the domain of physics include: an awareness of how passive sentences express meaning; an ability to interpret conditionals and other relationships as they are expressed in English; and an ability to interpret comparative language structures using "more," "-er," and "-est." Students also need to have a solid grasp of technical and related vocabulary that will be used on exams.

Teaching about Testing

While most professionals agree on the importance of teaching students content and language, there is some controversy about practicing test-taking and systematically exposing students to item-models in order to prepare students for the tasks they will face on state exams. LaSasso (1999) encourages the inclusion of test taking skills in deaf students' curricula and numerous schools spend considerable time preparing students for state tests. Johnson (2001a) also notes the need for preparing deaf and hard of hearing students for standardized tests, but warns that devoting too much time may hijack attention and energy needed for important social and linguistic

challenges unique to deaf individuals. Professionals such as Marshal (2001) criticize such practices, suggesting instead that classroom time be devoted only to high-quality aligned instruction rather than spent on boring, decontextualized, lower-level, skill-based test preparation. While development of content knowledge/skills should remain the primary intent of education, if students are to do well on written evaluations, they must also develop the ability to demonstrate their knowledge/skills.

Recognizing Limitations of Assessment

Finally, it is essential that any single evaluation not serve as the definitive description of a students' knowledge/skill competence. Attainment of a given score must not become the primary goal of instruction nor should failure to achieve a level of proficiency on a given evaluation measure necessarily be a mark indicating failure in a particular domain. Educators need to remain cognizant that evaluation results merely reflect student performance on a tool. If the tool is reliable, results may help to make decisions regarding student competence in the assessed area. However, tests in general contain random measurement error, and language-based tests given to deaf students in particular, may contain systematic bias. These testing characteristics may prohibit equating a student's obtained score with the student's "true score" (lack of reliability) and actual ability (lack of validity). All tools have limited usefulness in the decision making process, but tools that provide data lacking strong validity for their intended purpose,

should not be used, or at least not used alone, in assessing a student's competence. The use of data from such a tool will lead to invalid conclusions and possibly harmful decisions.

When making a determination about a student's content competence, the assessment process should consist of multiple evaluation measures and should take into account the strengths, weaknesses, and reliability of the evaluation tools used as well as the usefulness of their results for the purpose of drawing valid conclusions. By using multiple measures, professionals may have greater confidence in their ability to assess a student's actual ability in a given domain. This study has examined several of the features found in standardized tests currently in use by states to assess students' physics mastery. While these tests provide some data about student performance, they do not provide sufficient information to draw definitive conclusions about a deaf student's proficiency in the content area of physics.

Professionals' use of assessment can help or harm students. Johnson (2001a) warns "if used inappropriately, as appears to be the case in some states, they can become the single measure to earning a high school diploma—and this could have devastating impact on the academic and employment prospects of deaf and hard of hearing students. . . [they are being] placed in positions of vulnerability unparalleled by those of the general school population" (p. 1). We must strive to develop evaluation tools and assessment procedures that separate, as much as possible, language access

issues from content knowledge/skill. We must ensure students are given "adequate opportunity to receive reliable and accurate feedback about the range and extent of their academic progress in content areas" (Mitchell and Young, 2004 p. 26). Finally, we must recognize that assessment is fallible; as professionals we must do all we can to ensure that students are given the credit they deserve.

Center for Cognitive Science and Educational Practice. (2005). Readability Formulas and *Retrieved March 15,2005 from* http://csep.psyc.memphis.edu/cohmetrix/readabilityresearch.htm

Chal, J. S. & Dale, E. (1995). *Readability revisited: the new Dale-Chal Readability Formula.* Cambridge MA: Brookline Books.

Chiang-Soong, B., et. al., (1993). Readability levels of the science textbooks most used in secondary schools. *School Science and Mathematics, 93,* (1), 24-27.

Clinton, C (1999). *Contemporary assessment for educators.* New York: Longman.

Commonwealth of Massachusetts Department of Education (2004). MCAS Introductory Physics, Grade 9/10 Pilot Test.

Ewoldt, C. (1983). Test Simplification: a solution with many problems. *Perspectives in Education and Deafness, May/June,* 5-7.

Freile, V. E. (2003). Brighton writes own test. *Democrat and Chronicle.* Saturday October 25, 2003. Section B.

Hoff, D. J. (2004). Accountability Conflicts Vex Schools. *Education Week.* March 10, 2004, 23.

Homan, S, Hewitt, M. & Linder J. (1994). The development and validation of a formula for measuring single-sentence test item readability. *Journal of Educational Measurement, 31 (4),* 349-358.

International Association for the Evaluation of Educational Achievement (2003). *Trends in International Mathematics and Science Study. Retrieved February 3, 2005 from* http://timss.bc.edu/timss2003.html.

Isralelite, N. K. (1988a). On Readability formulas: A critical analysis for teachers of the Deaf. *American Annals of the Deaf. 133* (1), 14-16.

Isralelite, N. K. & Helfriech M. A. (1988b). Improving text coherence in basal readers: effects of revisions on the comprehension of Hearing-impaired and normal =hearing readers. *Volta review, 90,* 261-276.

Johnson, R. C. (2001a). High stakes testing and deaf students: some research perspectives. *Odyssey, 2* (3), 1.

Johnson, R. C. (2001b). High stakes testing and deaf students: some research perspectives. *Research Gallaudet. spring/summer 2001*, 1-6.

Kadamus, J. A. representing EMSC-VESID committee (2004). *Review of Other States' Approaches to Using State Assessments to Meet Graduation Requirements.* Attachment B of February 27, 2004 Memorandum to the Board of Regents. Retrieved April 19, 2005 from http://www.regents.nysed.gov/2004Meetings/March2004/0304emscvesidd2.htm

Kelly, R. R., Lang, H. G., Mousley, K., & Davis, S. M. (2003). Deaf college students' comprehension of relational language in arithmetic compare problems. Journal of Deaf Studies & Deaf Education, 8 (2), 120-132.

LaSasso, C. J. (1999). Test-taking skills: a missing component of deaf students' curriculum. *American Annals of the Deaf, 144 (1),* 34-43.

Massachusetts Department of Education (2001). *Massachusetts science and technology/engineering curriculum framework.*

Massachusetts Department of Education (2004). *Massachusetts comprehensive assessment system: Introductory Physics, Grade 9/10 pilot test.*

Martin, D. S. (2005). *Multiple-Choice Tests: Issues for Deaf Test Takers.* National Task Force on Equity in Testing Deaf Individuals Retrieved February 9, 2005, from http://gri.gallaudet.edu/TestEquity/mctests.html

Martin, D. S. & Mounty, J. L. (2003). *National Task Force on Equity in Testing: Deaf & Hard of Hearing Individuals.* Retrieved February 8, 2005, from http://gri.gallaudet.edu/TestEquity/lictests.html

McKey, B. G. and Lang, H. G. (1982). A comparison of deaf student's performance on true-false and multiple-choice items. *American Annals of the Deaf, 128* (1), 49-54.

Marshall, K. (2001). Test prep- the junk food of Education. *2001 Editorial Projects in Education, 23* (5), 30, 34.

Marshark, M., Lang, H.G., and Albertini, J.A. (2002). *Educating deaf students: from research to practice.* New York: Oxford University Press.

Moores, D. F. (2004a). High stakes: are the stakes too high? *American Annals of the Deaf, 145* (3), 235-236.

Moores, D. F. (2004b). No Child Left Behind: the good, the bad, and the ugly. *American Annals of the Deaf, 148 (5),* 347-348.

Murphy, P. (1996). The IEA assessment of science achievement. *Assessment in education: principles, policy, & practice, 3 (3),* 213-233.

Mitchell, R. E. and Young, T. A. (2004). Do Mathematics Test Scores Depend on Item Readability for Deaf and Hard-of-Hearing Students? Paper presented at the annual meeting of the American Education Research Association, San Diego, California April 13, 2004.

National Committee on Science Education Standards and Assessment, National Research Council (1996). *National Science Education Standards.* Washington, DC: National Academy Press. [Electronic version].

Orlich, D. C. (2004.) No Child Left Behind: An Illogical Accountability Model. *The Clearing House, 78 (1),* 6-11.Park, Chung; Allen, Nancy L. (1994). Relationships between Test Specifications, Item Responses, Task Demands, and Item Attributes in a Large-Scale Science Assessment. Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994). *ED371010.*

Paige, (2003). *No child left behind, final regulations: assessing students with disabilities.* Effective January 9, 2004. Retrieved March 11, 2005 from http://www.ed.gov/

Pfeffenberer, W., Zolandz, A. M., Jones, L. (1991). Testing physics achievement: trends over time and place. *Physics Today, Vol. 44 (9),* 30-37.Pratt, H. (2005). Where are we now? Two international studies help us assess science education reform efforts. *The Science Teacher, 72 (1),* 10-11.

Randall, K., McAnally, P., Rittenhouse, B., Russell, D., Sorensen, G. (2000). High Stakes Testing: What is at Stake? *American Annals of the Deaf, 145*(5), 390-393.

Resnick, L., Ed. (2003). Standards and tests: keeping them aligned. *Research points, 1,* (1), 1-4.

Rivera, C. & Standfield, C. W. (2003). The effect of linguistic simplification of science test items on score comparability. *Educational Assessment (9)*, 3/4, 79-105.

Rudner, L. M. (1978). Using standard tests with the hearing impaired: the problem of item bias. *Volta review, 80,* 31-40.

Sink, W. (2001). Evaluating a strategy for improving student performance on statewide standardized test. Capstone project: Montana State University.

Taylor, D. (2004). Intuitive Systems. Software program. Retrieved April 20, 2005, from http://www.readability.info

Traxler, C. (2003). *What is the reading level of deaf and hard of hearing people?* Retrieved May 15, 2005 from Gallaudet University, Research Institute Web site: http://gri.gallaudet.edu/Literacy/#reading.

Traxler, C. B. (2000). The Stanford achievement test, 9th edition: national norming and performance standards for deaf and hard of hearing students. *Journal of Deaf Studies and Deaf Education, (5)* 4, 337-348.

Trybus, R., Buchanan, C. & DiFrancesca, S. (1973). *Studies in achievement testing, hearing impaired students, United States: Spring 1971.* Series D, Number 13. Washington, D.C.: Office of Demographic Studies, Gallaudet College.

Trybus R. J. & Karchmer, M. A. (1977). School achievement scores of hearing impaired children: National data on achievement status and growth patterns. *American Annals of the Deaf 122* (1), 62-69.

University of the State of New York; State Education Department (1996). *Learning standards for mathematics, science and technology.*

University of the State of New York (2004). *Regents High School Examination: Physical Setting: Physics.* Wednesday, June 16, 2004 version.

Veronesi, P. (2000) Testing and assessment in science education: Looking past the scoreboard . *The Clearing House, 74 (1),* 27-30.

Yarroch, W. L. (1991). The implications of content versus validity on science tests. *Journal of Research in science teaching (28),* 7, 61-99-629.